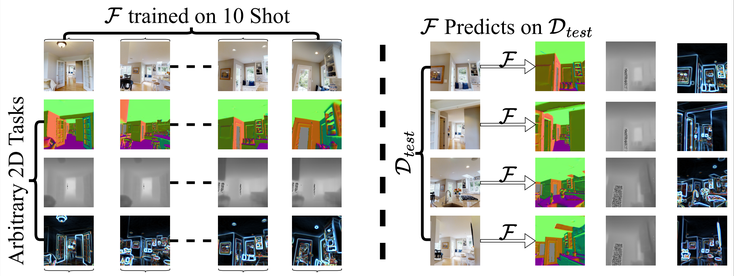
Universal Few -Shot Dense Visual Task Learner with Coupled Prompts

Jiho Choi under the supervision of Professor Hae-Gon Jeon

School of EECS, GIST, Korea

Promblem Definition



e.g., Semantic Segmentation, Surface Normal, Monocular Depth

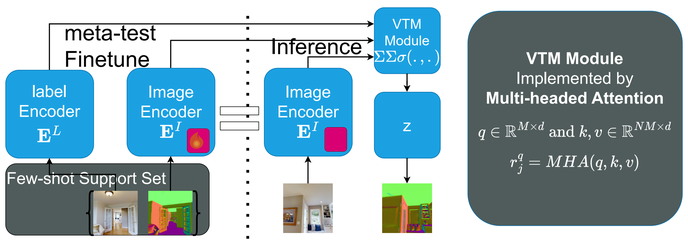
Previous Works

●VTM(ICRL 2023 oral)

$$y_j^q = \mathbf{D}^L(r_j^q), r_j^q = \sum_{i \le N} \sum_{k \le M} \sigma\left(\mathbf{E}^I(\mathbf{x}_j^q), \mathbf{E}^I(\mathbf{x}_k^i)\right)\mathbf{E}^L(\mathbf{y}_k^i)$$

Non-parametric token matching

to achieve universal few shot learner for dense prediction tasks

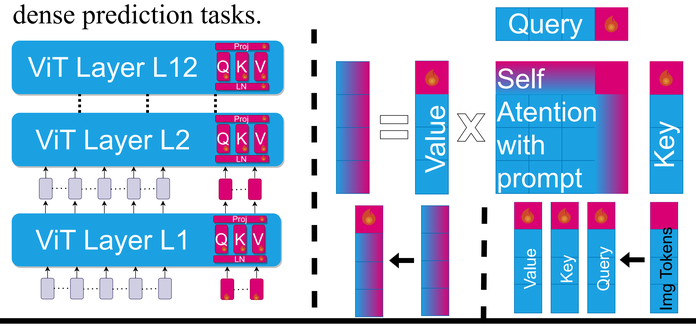


Contributions

- Devises novel method(**Weight Ensemble**) that can stabilize finetune on few-shot setup.
- Conducts extensive experiments to show effectiveness of **Prompt Tuning** on dense prediction tasks.
- Empirically shows that **knowledge** embedded in ViT's weight affects prediction performance significantly.

Prompt Tuning

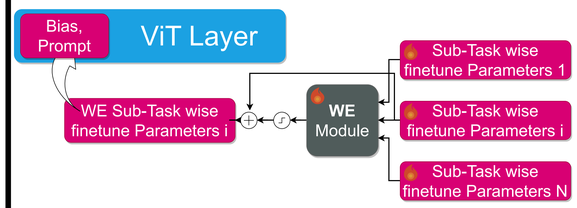
- Acts as task specific instruction that steers ViT based Encoders' pretrained knowledge.
- Shallow Prompt with Layer-wise Residual Prompts work for dense prediction tasks.



Methods

Weight Ensemble

- Handles Multi Channel Information
  e.g., surface normal(3 channel)
- Works as Attenuator for single channel tasks
  e.g., semantic segmentaion



Weight Ensemble (WE) module

$$\hat{P}_i := \sum_{j=1}^{T} \frac{P_j \omega_{i,j}}{\sqrt{\sum_{k=1}^{T}(\omega_{i,k})^2}}$$

$$W := (\omega_{i,j}) \in \mathbb{R}^{\mathbf{C}_\mathcal{T} \times \mathbf{C}_\mathcal{T}}$$

with
1. Lazy Update
2. Init Freeze
for more Stability

Pretrained knowledge

- Required to solve each specific task is quite different task-wisely
- Can Vision Foundation model be adopted as encoders without **catastrophic forgetting**?

ImagNet-22k
15M images
22K classes

Supervision on class prediction with cross entorpy loss

Segment Anything(SA-1B)
1+ Billion masks
11 million images

Supervision on mask prediction with focal loss and dice loss

Results

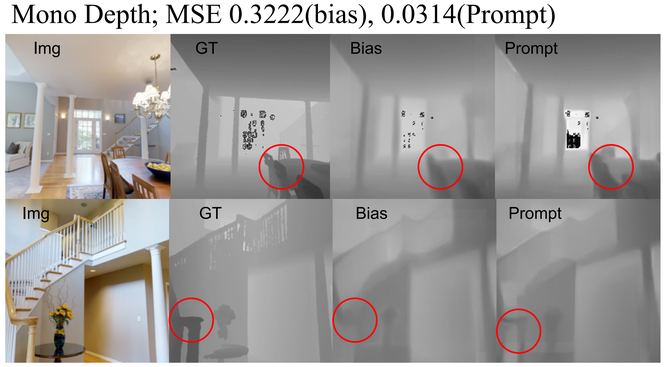
- Prompt, Weight Ensemble Performances on eash task

Supervision	Model	Tasks									
		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
		SS mIoU ↑	SN mErr ↓	ED RMSE ↓	ZD RMSE ↓	TE RMSE ↓	OE RMSE ↓	K2 RMSE ↓	K3 RMSE ↓	RS RMSE ↓	PC RMSE ↓
10-Shot	Paper	**0.4097**	11.44	**0.0741**	**0.0316**	**0.0791**	**0.0912**	0.0639	0.0519	**0.1089**	0.0420
	Reproduce	0.3875	10.59	0.08998	0.0322	0.08729	0.09897	0.06462	0.04853	0.1096	0.04229
	Prompts(Ours)	0.3820	**10.14**	0.0902	0.0358	0.0792	0.1019	**0.0594**	0.0491	0.1219	0.0435
	Weight Ensemble(Ours)	0.3032	10.82	0.08362	0.03286	0.08996	0.09698	0.06275	**0.04686**	0.11049	**0.04113**

- Performances on eash task for different pretrained weight

Supervision	Model	Tasks									
		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
		SS mIoU ↑	SN mErr ↓	ED RMSE ↓	ZD RMSE ↓	TE RMSE ↓	OE RMSE ↓	K2 RMSE ↓	K3 RMSE ↓	RS RMSE ↓	PC RMSE ↓
10-Shot	Paper	0.4097	11.44	**0.0741**	**0.0316**	**0.0791**	**0.0912**	0.0639	0.0519	**0.1089**	0.0420
	Reproduce	0.3875	10.59	0.08998	0.0322	0.08729	0.09897	0.06462	0.04853	0.1096	0.04229
	scratch	0.2784	**10.544**	0.08559	0.03344	0.08521	0.09034	0.06271	**0.04671**	0.1127	**0.04090**
	IN-22k	0.3401	10.613	0.08979	0.03386	0.08167	0.09350	**0.06112**	0.04860	0.1196	0.042190
	SAM	**0.4598**	-	-	-	-	-	-	-	-	-

Mono Depth; MSE 0.3222(bias), 0.0314(Prompt)



Surface Normal; mErr 10.59(bias), 10.14(Prompt)